虚无假设检验与 p 值计算的逻辑缺陷

姜红丙*　高琳　向玉琳　张晨熙

（郑州大学管理学院 郑州 450001）

摘要：科学研究中对虚无假设检验（NHST：null hypothesis significance testing）以及 p 值的误用、滥用已经相当严重。NHST 是 Fisher 显著性检验和 N-P 式假设检验的杂合体，但它又是如何杂合的，在计算步骤上如何体现？NHST 和 p 值计算的逻辑缺陷在哪里？这些问题并没有详尽的、通俗的解答。明确地阐述 Fisher 显著性检验、N-P 式假设检验、NHST 的步骤并加以分析和比较，辅之以典型示例进行 NHST 和 p 值计算的逻辑缺陷分析，能够给未在统计学领域深耕的经验研究者提供一定的启发。

关键词：p 值　虚无假设检验　Neyman-Pearson 假设检验　Fisher 显著性检验

中图分类号：C81　　**文献标识码**：A

Analysis of Logical Defects in Null Hypothesis Significance Test and p Value Calculation

Jiang Hongbing　Gao Lin　Xiang Yulin　Zhang Chenxi

（School of Management，Zhengzhou University，Zhengzhou，450001，China）

Abstract：The misuse and abuse of NHST（null hypothesis significance test）and p value are quite serious in scientific research. NHST is a hybrid of Fisher's significance test and N-P hypothesis test. But how is it mingled and how is it reflected in the calculation steps? Where are the logical flaws in NHST and p value calculations? There are no detailed and simple answers to these questions. Expounding, analyzing and comparing the steps of Fisher's significance test, N-P hypothesis test and NHST clearly, combined with a typical example for the logical defect analysis of NHST and p-value calculations, can provide some inspiration for the empirical researchers who are not deeply involved in the field of statistics.

Key words：p value　NHST null hypothesis significance testing　Neyman-Pearson hypothesis test　Fisher significance test

1　引言

经验研究中占主导地位的研究策略，假说演绎法（hypothetico-deductive method），一般从研究问题出发，综述相关文献和理论，讨论有助于回答研究问题的理论假设，进而从理论假设中推导出研究假设（hypothesis）；接着进行研究设计（选取或开发构念测量工具，设计数据收集方法和数据分析方法）；之后进行实际数据收集、预处理、数据分析、解读；最后进行总结，撰写报告[1]。其中，一个核心环节是虚无假设检验[2，3]。一般而言，研究假设作为备择假设出现在虚无假设检验中，在虚无假设检验过程中研究者希望 p 值小于某个规定的数值（如 0. 05、0. 001 等），以拒绝虚无假设，支持研究假设。但由于研究者错误地将 p 值作为代表证据的强度，在追求论文发表的过程中，他们往往会忽视论文内容的真实性，从而导致发表偏倚。因此，一些期刊明确要求禁止将 p 值作为衡量

* 本文系国家自然科学基金项目（项目编号：71801195）与河南省软科学研究项目（项目编号：242400410065）的研究成果之一。通讯作者：姜红丙，博士，现为郑州大学管理学院副教授，副院长，研究方向为管理研究方法论、创新方法论。Email: jhbymx@zzu.edu.cn

研究合理性的唯一标准。例如，《美国公共健康杂志》（*AJPF*）从 1983 年起就要求投稿者删除所有 p 值，否则就请转投其他杂志；《流行病学》（*Epidemiology*）在 1990 年创刊之初也公开声明："作者向本刊投稿时，若忽略显著性检验，将有助于提高稿件被录用的可能性……我们根本就不采用这一方法"[4]。2016 年，美国统计学会声明了关于 p 值的 6 个原则[5]。虽然这 6 个原则是统计学家的老生常谈，但是，这是一个国际上极具影响力的统计学组织第一次为 p 值的问题发表声明。不久之后，美国政治学顶级学术期刊《政治分析》（*Political Analysis*）在 2018 年发表声明，以"p 值本身无法提供支持相关模式或假说之证据"为由宣布禁用 p 值。可见，科学研究中对虚无假设检验以及 p 值的误用、滥用已经相当严重了[6]。

吕小康（2014）[4]认为 NHST 是 Fisher 显著性检验和 Neyman-Pearson（以下称二人为 N-P）式假设检验的杂合体，它既不完全是 Fisher 显著性检验，也不完全是 N-P 式假设检验，它是杂合实用性与数学之美的折衷体现。但是 Fisher 显著性检验和 N-P 式假设检验是怎么杂合的？这种杂合在计算步骤上是如何体现的？吕小康（2014）对此未作详尽的回答。另外，在进行虚无假设检验的过程中，p 值的应用也备受争议，究竟是 p 值被我们误用了，还是其计算本身就存在逻辑缺陷？郝丽等（2016）[7]的观点是 p 值被误用了，其误用的原因，一是，将 p 值简化为"p 值是原假设为真的概率"，即在 NHST 中，p 值与 N-P 式假设检验中的α值是"等价的"[4]。二是，大多研究都是先"按'原假设为真'推断至'备择假设为假'，再将'p 值是原假设为真的概率'引申到'p 值是备择假设为假的概率'"，也就是说，p 值本身只是用于测量原假设的证据，但是大多学者认为 p 值可提供足够的证据对研究假设（备择假设）进行判断。另外一些学者则认为 p 值计算本身也是有逻辑缺陷的，例如 Lindley（1993）[8]指出由于设想的试验方案不同，p 值的计算结果也可能不同。

综上所述，Fisher 显著性检验与 N-P 式假设检验的不合理杂合造成了 NHST 的逻辑缺陷，这也是 p 值被误读或误用的原因，另外 p 值本身也存在计算上的逻辑缺陷。本文采用文献分析法，通过对 Fisher 显著性检验、N-P 式假设检验和 NHST 虚无假设检验三者的检验步骤进行比较，详细地分析了 NHST 是前两者如何杂合而成的。同时，通过具体的算例来分析 NHST 和 p 值计算的逻辑缺陷。接着，回答了为什么 NHST 和 p 值虽屡受批评，却仍能大行其道。本文可能的贡献不在于创造了关于假设检验的更多新知识，而在于澄清问题，为没有在统计学深耕的经验研究者提供通俗的解释，使他们更容易理解 NHST 和 p 值计算的逻辑缺陷。

2　NHST 是 Fisher 显著性检验和 N-P 式假设检验的杂合体

2.1　Fisher 显著性检验

实际上，NHST 是 Fisher 和 N-P 关于假设检验思想的杂合体，内部存在着种种矛盾[9, 10]。Fisher 模式下的 p 值是在某个假设 H0 为真，其他相关假设也为真的前提下[6]，试验数据出现当前或更加极端值的概率。p 值小于给定的某个值，仅仅表明这个假设 H0 是错的，或者小概率事件发生了[11]，但是通常无法根据一次试验，确定 H0 被拒绝或者不被拒绝。Fisher 认为当满足下述条件时才能合理地拒绝假设 H0：在试验设计没有重要错误的前提下，进行多次试验，这些试验结果中，统计意义显著的结果在数量上具有压倒性优势[12]。所以，一次试验得出了统计意义显著的结果，仅仅能给我们一些提示性的证据，说明试验结果值得

我们注意，需要进行更深入的研究。如果要验证另外一个假设，则需设计另外一套检验程序，而不是在一次检验中拒绝某一假设、同时接受另外一个假设。因此，Fisher 认为备择假设的引入完全是没有必要的[13]。

Fisher 显著性检验步骤总结如下[3, 14, 15]：

（1） 确定统计假设 H0；

（2） 选择合适的检验统计量 T，确定其在 H0 为真的前提下的分布；

（3） 根据当前的试验数据计算检验统计量 T 的数值 t；

（4） 根据 T 的分布（在 H0 为真的前提下的分布），确定与 t 相对应的显著性水平 p；

（5） 如果获得的 p 小于预期的值，则要么 H0 不为真，要么小概率事件发生了。

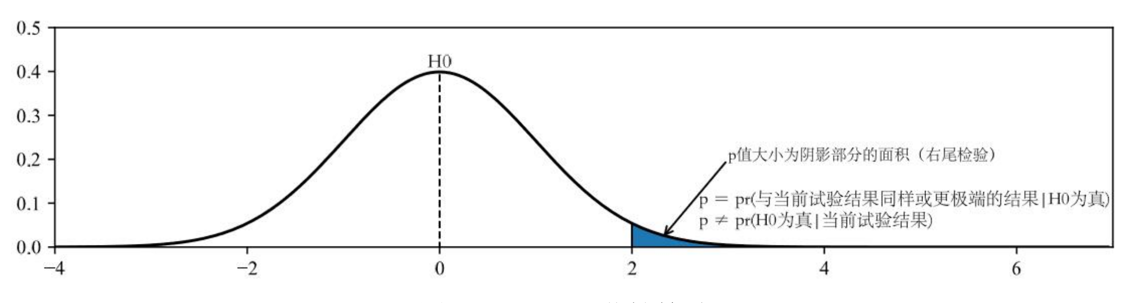


图 1　Fisher 显著性检验

2. 2　Neyman-Pearson 式假设检验

N-P 式假设检验则采用虚无假设（null hypothesis）对备择假设（alternative hypothesis）的形式。检验的要旨为在限制第一类错误的概率不超过显著性水平α的条件下，谋求第二类错误的概率β的最小化[16]。N-P 式的假设检验思想是基于重复抽样的前提得出的，并不能保证根据某一个样本的观测结果所做出的接受或拒绝的决策是对还是错[13]。同时，在 N-P 的思想框架中，完全没有提到 p 值，他们使用拒绝域来对假设进行判断。

N-P 式假设检验缘起于工业生产中质量控制的需要[17]。例如：一个生产螺丝帽的工厂接到一个订单，要求生产的螺丝帽直径为 2 ± 0. 01cm。此时，我们就会很清楚，要想检测出质量有问题的螺丝帽，所需的最小的效果量（effect size）是 0. 01cm。我们也很容易控制样本容量，因为从成千上万个螺丝帽中抽出容量较大的样本并不难。一方面，如果工厂无法有效检测出直径过大（ ＞ 2. 001cm）或过小（ ＜ 1. 999cm）的螺丝帽，也就是犯第二类错误的概率太大，可能会导致订单被取消。另一方面，如果实际上符合规定直径标准（2 ± 0. 01cm）的螺丝帽，经常被检测为不合格产品，也就是犯第一类错误的概率太大，也会给工厂带来不少的损失，导致生产成本的无谓增加。因此，在这种情境下，我们可以把两类错误的减少或增大所带来的好处或坏处，通过适当的转换，用金钱来衡量。这也就意味着，我们可以通过数学优化的方法求出最恰当的α和β值，即最小化损失或最大化收益。

但是在大多数实际研究中，我们并不能轻易地控制样本容量、也不容易确定最小效果量以及合理地α和β值。例如，当我们要检验两组人智商是否有明显差异时，常常并不清楚所需的最小效果量是多少。尽管一些经验法则（rules of thumb）存在[18-20]，并告诉我们什么是大效果量、中等效果量和小效果量，但是判断效果量的大小极其依赖所研究的问题。例如研发一种新药，哪怕提升治愈率的效果量很小，也是可以接受的。

Neyman-Pearson 的假设检验步骤总结如下[15, 21, 22]：

（1） 确定两个统计假设①：虚无假设 HM 与备择假设 HA；

（2） 选择合适的检验统计量 T，确定其在 HM 为真的前提下的分布；

（3） 指定能够接受的犯第一类错误的最大概率α；

（4） 根据（1）、（2）、（3）和指定的统计功效②、最小效果量等，计算最小的样本容量；

（5） 根据 Neyman-Pearson 引理及其扩展定理，计算拒绝域 C；

（6） 根据当前的试验数据计算检验统计量 T 的数值 t；

（7） 如果 t 在拒绝域 C 中，则拒绝虚无假设，接受备择假设，否则接受虚无假设。

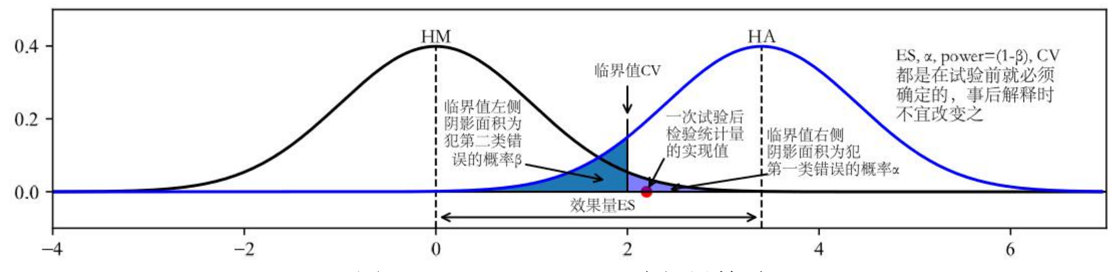


图 2　Neyman-Pearson 式假设检验

2.3　虚无假设检验（NHST）

虚无假设检验是 Fisher 显著性检验和 N-P 式假设检验思想的杂合，其典型的检验步骤如下[4, 15, 23]：

（1） 确定两个统计假设：虚无假设 HM 与备择假设 HA；

（2） 选择合适的检验统计量 T，确定其在 HM 为真的前提下的分布；

（3） 指定能够接受的犯第一类错误的最大概率α；

（4） 根据当前的试验数据计算检验统计量 T 的数值 t；

（5） 根据 T 的分布，确定与 t 相对应的显著性水平 p；

（6） 若 $p <= \alpha$，拒绝 HA，接受 HM；若 $p > \alpha$，接受 HA，拒绝 HM。

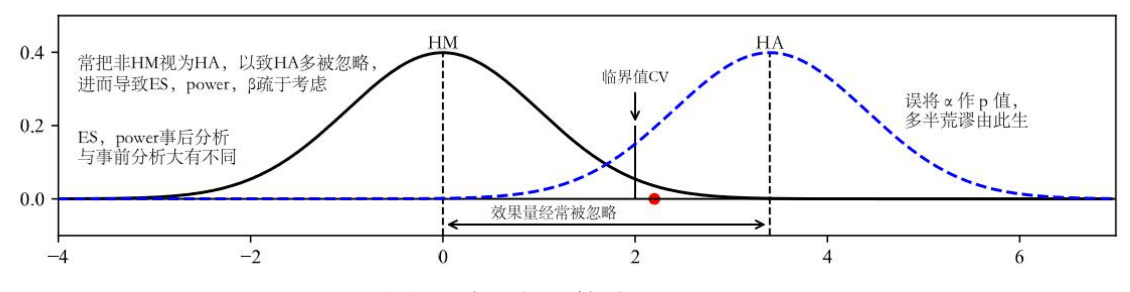


图 3　虚无假设检验（NHST）

①这里的假设是指简单假设（simple hypothesis）。如果一个统计假设能完全确定总体的分布，则称此假设为简单统计假设；否则称之为复合统计假设（composite hypothesis）。在虚无假设和备择假设都是简单假设的情况下，利用尼曼—皮尔森引理（Neyman–Pearson lemma）可以确定最大功效检验（most powerful test）的形式。一般来讲，检验 H_0 vs. H_1 的最大功效拒绝域（best critical region）与检验 H_0 vs. H_2 的最大功效拒绝域不一定相同。所以当备择假设为复合统计假设的情况下（例如 H_0：$\mu = 0$ vs. H_1：$\mu>0$），如何确定其最大功效拒绝域，Neyman-Pearson 并没有完全解决这个问题。卡林—鲁宾对尼曼—皮尔森引理进行了拓展，提出了卡林—鲁宾定理（Karlin–Rubin theorem），通过该定理，可以导出某些复合假设检验问题的一致最大功效检验（uniformly most powerful test）。

②备择假设是复合假设时，统计功效的计算就很复杂了。假如这个复合假设包含如下简单假设：H_1，H_2，...，H_n，...那么对于其中每一个简单假设都会有一个统计功效，而且一般不相同，这些功效可能是：$1-\beta$（$H_1|\alpha$），$1-\beta$（$H_2|\alpha$），...，$1-\beta$（$H_n|\alpha$），...此时，对于这个备择假设为复合假设的检验来说，称功效函数（power function）比称功效（power）更合适。功效和效果量都可以事先指定一个标准水平，然后根据样本计算实际水平。

NHST 的杂合主要体现在，NHST 的检验步骤中，步骤（1）、（2）、（3）、（4）与 N-P 式假设检验中步骤（1）、（2）、（3）、（6）相同，步骤（4）、（5）与 Fisher 显著性检验中步骤（3）、（4）相同，而在步骤（6）中，NHST 不存在 N-P 式假设检验中统计功效、最小效果量及拒绝域等相关信息，错误地认为 Fisher 显著性检验中的 p 值与 N-P 式假设检验中的α值是“等价的”，无论是 Fisher 还是 Neyman-Pearson 都不会认同 NHST 的计算过程。直观看上去，Fisher 显著性检验的假设 H0 与 N-P 式假设检验下的虚无假设 H_M 没有区别，仅仅是后者多了一个备择假设 H_A。实际上两种模式有很大不同。我们只要把 N-P 式假设检验下的统计假设改变一下形式就很容易看出来二者的不同[22]。H_M：$M_1 - M_2 = 0 \pm$ MES（最小效果量）；H_A：$M_1 - M_2 \neq 0 \pm$ MES。M_1 是由 H_M 确定的概率分布的参数，M_2 是由 H_A 确定的概率分布的参数。如果研究设计中没有用到备择假设 H_A 提供的信息（即最小效果量和犯第二类错误的概率），那么 N-P 式的假设检验就退化到 Fisher 显著性检验模式。常见的统计软件如 SPSS 是以 Fisher 的统计检验思想为主的[22]，这就意味着在大多数时候，我们读到的经验研究论文中所使用的都是 Fisher 显著性检验，进一步说，作为假设检验中备择假设的研究假设基本上就是摆设。

综上，NHST 分别利用了 Fisher 模式和 N-P 模式中对我们有吸引力的地方，而忽视了这两种模式发挥作用的前提条件。具体而言，NHST 仅仅利用了 Fisher 模式下 p 值能够测量试验数据对某个假设 H0 的支持程度这一便利之处和 N-P 模式易于决策的优点。但是忽视了，如果要做出有效的决策，在 N-P 式假设检验下需要考虑更多的信息，如β值、功效值、最小效果量、备择假设 H_A 等信息，以至于在实际操作中误认为“非 H_M”就等同于备择假设 H_A [24-26]。

3　虚无假设检验的逻辑缺陷分析

3.1　经常忽略备择假设提供的信息

社会科学的方向性假设没有用到备择假设提供的信息，因此本质上是一种 Fisher 显著性检验。只要拒绝了虚无假设，不管备择假设是什么最终都会被接受。这是对统计的极大地误用。被记者 Geoffrey Wansell 称为“现代英国法律史上最大不公”的 Sally Clark 案最有力地说明了这一点。一个名叫 Sally Clark 的妇女第一个孩子在出生后不久离奇死亡，医生查不出其它病因，结果诊断为一种 SIDS（婴儿猝死综合症）的病因。不幸的是，Sally Clark 的第二个孩子在出生后不久也去世了，警方就怀疑是 Sally Clark 杀死了自己的两个孩子。儿科专家 Roy Meadow 用统计证据和推理说服了陪审团，结果 Sally Clark 以谋杀的罪名被判处终身监禁。Roy Meadow 的理由是，在 Clark 这样的家庭中，一个婴儿患 SIDS 的概率仅为 1/8543，两个婴儿患 SIDS 的概率为 $1/8543^2 \approx 1/73000000$；因此 Clark 无罪的概率仅为 1/73000000；小概率事件发生了，所以 Clark 有罪。用 NHST 的语言表达：H_M：两个婴儿死于 SIDS；H_A：Sally Clark 杀死了两个婴儿。两个婴儿死于 SIDS 的概率太小，于是拒绝 H_M，拒绝了虚无假设意味着接受备择假设 H_A，因此 Clark 有罪。在这样的推理中，H_A 只是摆设，完全没有起到作用。事实上，一位母亲杀死亲生婴儿的概率比 SIDS 杀死婴儿的概率更低。其中一种说法是母亲杀死亲生婴儿的概率约为 1/92000[27]，比死于 SIDS 的概率 1/8543 更小。也就是说，在上述例子中，如果只能在 H_A 与 H_M 中做选择的话，理性的选择应该是 H_A：两个婴儿死于 SIDS。

3.2 混淆 p 值与α值的含义

如果说容易致人忽略备择假设提供的信息不是 NHST 本身存在的问题，那么混淆 p 值与α值的含义，NHST 难辞其咎[28]。p 值是 Fisher 提出来的，是在某个假设 H 为真，其他相关假定也为真的前提下，试验数据出现当前值或更加极端值 D 的概率[15]，记为 pr（D|H）。p 值是当前试验数据的性质，测量的是当前数据及更加极端的数据反对假设 H 的程度。每一次试验都有一个 p 值，也就是说，p 值是一个随机变量，进行一次试验就会计算一个 p 值。α值是 Neyman−Pearson 提出来的，其表达的含义是在虚无假设 H_M 正确的前提下，进行了 N 次（足够多）试验，其中 H_M 被拒绝的次数不超过 N×α次。α值是检验（test）的性质，并非试验数据的性质。p 值和α的区别主要在于：（1）p 值是一次试验的结果；α是 N 次试验的结果；（3）p 值是数据的性质，α是检验的性质；（2）p 值的阈值（如 0.05、0.001 等）可以在试验前确定，也可以在试验后确定；α的阈值必须在试验前确定，因为 N−P 式假设检验下拒绝域的计算需要用到α值；（4）可以根据 p 值大小得出数据对 H 支持程度的连续性强弱判断；α则不能，它只能做三分决策，要么支持，要么拒绝，或有待进一步考察。

混淆 p 与α的含义是 NHST 逻辑混乱的根源。p 值是一个随机变量，进行一次试验就会计算一个 p 值，p 值小于 0.05 还是 0.001 在试验前决定或在试验后决定没有区别；但是α不同，它是人为地预先确定的一个值，试验者正是根据这个值和其它设定的值，来设计试验的，设计试验的时候不管某一次试验是否结果显著，但是能保证长期来看犯第一类错误的概率一定小于α。用一个具体例子来说明：设定 Fisher 模式下的 p 值为 0.05，设定 N−P 模式下α阈值也为 0.05。然后进行 10000 次试验（假设已经足够多），在 Fisher 模式下 p 值小于 0.05 的试验次数并不确定，可能是 20、50、100、500 次等；但是在 N−P 模式下犯第一类错误的决策次数（根据概率理论）不会超过 25 次（假设 10000 次试验中 H_M 有 5000 次为真）。

3.3 得不到我们想要的结果

我们最想得到这样的结果：仅根据一次试验的结果就能确定，当试验结果显著时，原假设 H_0 或备择假设 H_1 为真的概率。但在实际计算过程中，我们往往得不到这种想要的结果。Fisher 假设检验中最吸引我们的是 p 值，在 NHST 虚无检验模式中等同于α值，我们通常错误地认为α是：如果一次试验结果显著，我们拒绝虚无假设所犯的概率。例如，假设一次研究结果得到 $p < \alpha = 0.05$，它常被解释为，如果拒绝了虚无假设，那么在 100 次试验中，我们错误的次数不会超过 5 次。这是我们十分渴望的结果，但是，事情远没有这么简单。做这样的判断得有个前提，那就是假设检验的虚无假设本身是真的。如果这个前提不成立，上述理解就不成立。如图 4 所示，虽然α = 0.05，125 个显著结果中可能 36%都是错误的，远大于我们认为的 5%的显著结果是错误的。实际研究中，虚无假设中真：假为 9:1 是很正常的[17]。

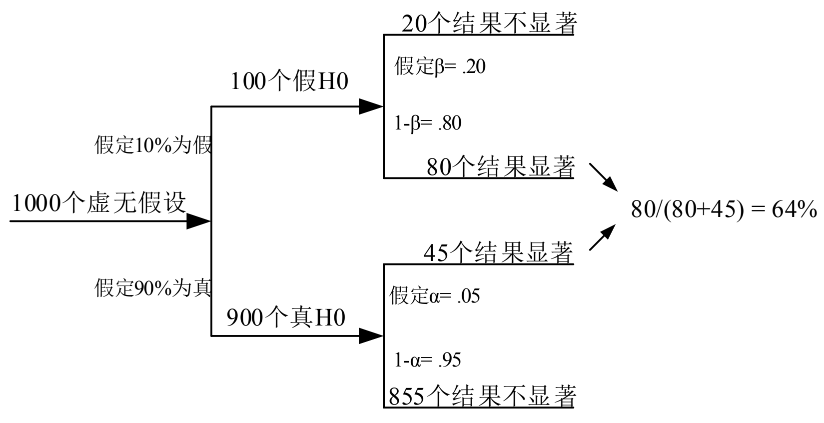


图 4　例释α的含义

还需要注意的是α、β都是长期试验条件下的概率，并不是一次试验条件下就能够得到的。什么叫长期试验？在满足概率公理化定义的前提下，频率学派对概率值的确定方法为[29]：在 N 次重复试验中，事件 A 发生 K_N 次，则事件 A 发生的频率为 P_N（A）= K_N / N = 事件 A 发生的次数/重复试验次数。长期试验与 N 的大小有关，随着重复次数 N 的增加，频率会稳定在某一常数附近。这个常数已经与 N 无关，它就是事件 A 发生的概率 P（A）。Fisher 和 Neyman，Pearson 都是频率学派的主要代表人物。Neyman 和 Pearson 关于假设检验的理论是建立在概率的频率解释基础上的[30]。但是在现实世界里，我们无法把一个试验无限次地重复下去，因此要获得 P（A）是很难的。我们有可能做到的常常是重复试验足够多次，用 P_N（A）去近似地代替概率 P（A）。

表 1　虚无假设检验（NHST）涉及的核心概念

概念的组合条件	概念								
	p	α	1−α	FRP	β	power	TRP	W_1	W_2
一次试验	√							√	√
长期试验		√	√	√	√	√	√		
pr(试验结果或更极端 | H_0 为真)	√								
pr(试验结果显著 | H_0 为真)		√							
pr(试验结果不显著 | H_0 为真)			√						
pr(试验结果显著 | H_1 为真)						√			
pr(试验结果不显著 | H_1 为真)					√				
pr(H_0 为真 | 试验结果显著)				√				√	
pr(H_1 为真 | 试验结果显著)							√		√

注：背景为灰色的概念 W_1、W_2 是我们最想要的，例如 W_1 指的是根据一次试验我们想知道 pr(H_0 为真|试验结果显著)，可惜 NHST 无法给我们这样的结果。但是现实中，我们发表的论文绝大多数都统计意义显著（否则评审通不过），统计意义不显著的结果很少。因此，我们也不知道到底有多少躺在抽屉里的统计意义不显著的研究结果。更为糟糕的是，在论文 GDP 的大环境下，有些研究者不希望辛辛苦苦做的结果躺在那里浪费了，转而采取一些有问题的行为使得研究结果显著。

4　p 值计算的逻辑缺陷分析

4.1　p 值是如何计算的

那是 20 世纪 20 年代后期，在英国剑桥一个夏日的午后，一群大学的绅士和他们的夫人们，还有来访者，正围坐在户外的桌旁，享用着下午茶，在品茶过程中，一位女士坚称把茶加进奶里，把奶加进茶里，这两种不同的做法，会使茶的味道品起来不同。这位女士的观点，激起了大家的兴趣，有人说让我们来检验这个命题吧。Fisher 在 *The Design of Experiments* 的第二章详细地介绍了如何设计各种不同的方案来检验这个女士的命题[12]。但是 Fisher 的介绍方式太过复杂，

这里我们使用经 Lindley 改进过的、又不失 Fisher 思想精髓的方案[8]来介绍 p 值到底是如何计算的？

改进后的试验设计为，每次呈现给女士两杯茶，并告诉她，其中一杯是先加奶后加茶的另外一杯是先加茶后加奶的，需要这位女士指出，到底哪一杯是先加茶，哪一杯是先加奶？如果女士的判断是正确的，记作 R，否则，记作 W。现在该试验重复了六次，假设试验结果为，RRRRRW，只有最后一次判断是错误的，那么 Fisher 的分析如下。

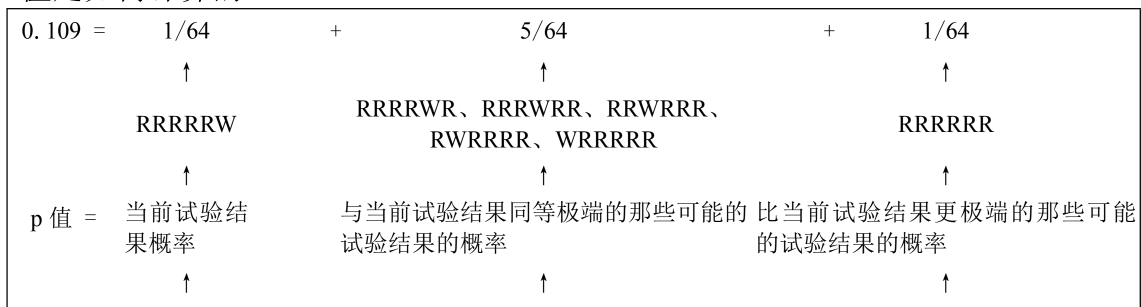
首先，假设这位女士根本就不能区别哪杯茶是先加奶后加茶，哪杯茶是先加茶后加奶的（原假设）。这也就意味着，女士的每一次判断都是随机的，每次判断正确的结果和错误的结果概率均为 1/2，并且每次试验与其余试验是相互独立的。那么观察到的试验结果，RRRRRW 的概率就为 $1/2^6 = 1/64$。Fisher 就做出判断：要么原假设是正确的，一个小概率事件发生了；要么原假设是错误的，也就是说，这位女士有能力区别哪杯茶是先加奶或者后加奶的。

试验的结果为 1/64，属于小概率事件。那么按照 Fisher 的分析，我们可能会倾向于认为原假设是错误的。Fisher 马上意识到了这样的分析是错误的，因为这样的试验安排，也就是 6 次的重复试验，任何一种可能结果的概率都是 1/64，所以由此而确定某个假设的对错明显是荒谬的。

为了避免这种荒谬，Fisher 声称只要试验结果为 5 个正确，1 个错误（不管这一个错误是在哪一次品茶的结果）都有同等的证据力度，W 可能出现在 6 个位置上的任意一个，那么概率就为 $6/2^6 = 6/64 = 0.094$，在 5%的显著性水平上，并不显著。于是避免了任何观察到的试验结果都显著这样的荒谬结果。

但是，Fisher 马上意识到了，这仍然不能解决问题。例如，让该女士做 300 次判断，其中 150 次正确，150 次错误的概率为：$C_{300}^{150}\times(1/2300) = 0.046$。这是连续做 300 次判断，最有可能发生的结果，其他任何一种结果的概率都比 0.046 小。于是同样的问题又出现了，任何一种结果都在 5%的显著性水平上显著，这仍然是不合理的。怎么解决这个问题呢？我们暂时回到 6 次重复试验的结果，RRRRRW。天才的 Fisher 想到如果 5 次正确 1 次错误，能够拒绝原假设的话，那么 6 次全对，肯定更能拒绝原假设，因此也要把 6 次全对的概率也加进去，0.094+1/64 =0.109。同样的道理用在 300 次的重复试验中，判断正确次数超过 150 次的那些试验结果的概率也应加入进来。于是 p 值应为：$(C_{300}^{150}+C_{300}^{151}+\cdots+C_{300}^{300})\times(1/2300) = 0.523$。

总结一下，p 值其实是由三部分概率构成的：（1）当前试验结果的概率；（2）与当前试验结果同等极端的那些可能的试验结果的概率；（3）比当前试验结果更极端的那些可能的试验结果的概率。图 5 以女士品茶试验为例，展示了 p 值是如何计算的。



观察到的	未观察到的	未观察到的

图 5　p 值的计算方法

4.2　p 值计算的逻辑缺陷

p 值到底是哪个事件的概率呢？实际上它不是一个事件，而是多个事件的概率之和。它涉及到当前的试验数据加上与当前数据同等极端和比当前数据更为极端的数据[31]。由图 5 我们可以看出，p 值由观察到部分的概率和未观察到部分的概率加总而成。p 值的逻辑问题主要就出现在未观察到的部分的概率怎么计算[8,32-34]。如何定义与当前试验结果同等极端的那些可能的试验结果？如何定义比当前试验结果更极端的那些可能的试验结果？p 值的逻辑困境就在于此。同样以女士品茶这个例子进行说明，这里假设我们有两种试验方案，第一种，我们前文已经介绍过了，那就是让该女士做 6 次判断，每 6 次判断的结果记录下来，记为一次试验结果；第二种试验方案为：连续不断地让该女士做判断，直到第一次错误判断出现，记录下来这些判断的结果就是一次试验结果。

假设我们现在得到的试验结果为 RRRRRW，接下来我们来计算 p 值。在第一种试验方案的框架下，前文我们已经分析过了，p 值等于 0.109，在 5%的显著性水平下，原假设是不能被拒绝的。第二种试验方案的框架下 p 值为多少呢？我们按照同样的思路进行计算：（1）当前试验结果的概率为 1/64；（2）与当前试验结果同等极端的试验结果的概率为 0；（3）比当前试验结果更极端的试验结果为 RRRRRRW、RRRRRRRW …，因此它们概率和为 $1/2^7 + \cdots + 1/2^\infty = (1/2^7) \div (1-1/2) = 1/64$。那么第二种试验方案框架下的 p 值为：$1/64 + 0 + 1/64 = 0.031$。在 5%的显著性水平下，原假设是要被拒绝的。这样矛盾就出现了，我们告诉了这位女士同样的信息（试验方案是我们头脑当中设想的，并没有告诉这位女士），这位女士诚实地进行了判断，相同的试验结果（RRRRRW）得到的推论却是矛盾的，这显然是不够合理的。

p 值的另外一个逻辑问题是，相同的 p 值是否意味着相同的证据力度？假设有相同的原假设，试验一的样本容量为 20，得到的 p 值为 0.042，试验二的样本容量为 100，得到的 p 值也为 0.042，那么试验一的结果与试验二的结果是否有相同的证据力度呢？如果证据力度不同，哪一个试验结果反对原假设的力度更强呢？学术界存在着极大的争议。一些学者认为，试验一的证据力度更强[35-37]，另外一些学者则认为试验二的证据力度更强[38]，还有一些学者则认为，试验一与试验二的证据力度相同，特别是 Fisher 本人就持这种观点：只要计算的方法正确，相同的 p 值就有相同的证据力度[39]。

5　为什么虚无假设检验和 p 值仍能大行其道

NHST 和 p 值虽然受到众多学者质疑和批判，但仍被广泛推崇的原因可归结为以下几点：

实用性：Fisher 作为一名有着丰富工作经验的应用统计工作者，深知统计工具的使用应该着重其在工作中的实用性，而 Neyman 和 Pearson 作为纯数理工作者，对于统计工具追求数学上的精确和完美，这也导致了两者之间出现严重的分歧。而作为二者杂合体的 NHST 吸收了 Fisher 模式中 p 值测量试验数据对某个假设 H 的支持程度和 N-P 式在决策上的便利性[40]。Yes or No 的决策是我们在科学研究和日常生活中回避不了的事实，科研人员对没有根据的 Yes or No 决策（特别是没有数字作为支撑的决策）抱有强烈的戒心，而 NHST 和 p 值极大地满足了人们的这种心理需求，因此它们有肥沃的生存土壤。

期刊导向：社会科学领域内顶尖期刊在论文发表上的榜样效应为 p 值的广泛使用推波助澜[①]。进而演化为一种标准化的实证研究程序和方法论要求，即凡是统计推论均须进行假设检验，而进行假设检验就要报告 p 值，这种要求又通过各学科内的统计教材反复示范，最终在整个学科领域全面制度化[5]。

科学性焦虑：根据吕小康（2014）[4]的观点，“社会科学中无论是哪门学科，都存在不同程度的这种压力，……它们对外仍需不断‘证明’自己是一门科学事业，对内还需整合学科体系和发展导向。解决这些压力的方式之一，就是建立一个整合的分析框架，确定整个学科的基本理论范式，同时引入一系列的数学工具”，而 NHST 和 p 值就在一定程度上担当了这样的角色。

6 研究结论

本文在吕小康（2014）[4]、郝丽等（2016）[7]的研究基础上，提出三个研究问题：（1）虚无假设检验（NHST）的逻辑缺陷在哪里？（2）p 值计算的逻辑缺陷在哪里？（3）为何二者有逻辑缺陷却仍能大行其道？

对于研究问题（1），本文认为 NHST 的逻辑缺陷关键在于混淆 p 值与α的含义，进而忽略备择假设提供的信息，导致无法获得我们想要的试验结果；对于研究问题（2），本文认为 p 值计算的逻辑缺陷之处在于：①根据相同的试验结果，如果设想不同的试验方案，计算所得的 p 值也可能不同；②在解读 p 值时，相同的 p 值是否意味着相同的证据力度，也没有定论；对于研究问题（3）的回答是：①它极大地降低了研究者对于不确定和没有数据的决策的戒备心；②顶尖期刊在论文发表上的榜样效应为 NHST 和 p 值的广泛使用推波助澜；③社会科学里的各个专业需向外界证明自己是一门科学事业，因此就需要建立一个整合的分析框架，而 NHST 和 p 值正好满足了这一需求。

本文并不否认 NHST 和 p 值在心理学、社会学等学科中的作用，这一点与吕小康（2014）[4]、姜红丙（2017）[1]等人在其论文中的观点一致。虽然 NHST 和 p 值屡受批评，但它仍然是学者进行经验研究的首选，更多地是因为研究者对研究工具使用的惯性，而改变这种惯性需要足够长的时间[13]。但是需要我们清醒认识到的是，无论是 Fisher 显著性检验还是 N-P 式假设检验，都需要长期试验才能得出结论。我们往往缺乏这样的耐心，总希望能够在一项研究中得到一个明确的结论。在这一目的上，NHST 和 p 值肯定让我们失望。但若因此而完全禁用它们也没有必要[41]，真正地了解 NHST 和 p 值计算的逻辑缺陷之后，审慎地使用它们，才是较为合理的取向。

[①]例如，1991 年《美国社会学评论》制订的一项新的发表要求明确规定禁止使用 0.05 以上的显著性水平，且必须使用“*”、“**”、“***” 分别表示 $p<0.05$、$p<0.01$、$p< 0.001$。

参考文献

[1]姜红丙．科学视角主义对管理研究的启示[J]．外国经济与管理，2017，39(03)：99-113.
[2]风笑天．社会学方法二十年：应用与研究[J]．社会学研究，2000(01)：1-11.
[3]焦璨，张敏强．迷失的边界：心理学虚无假设检验方法探究[J]．中国社会科学，2014(02)：148-163+207.
[4]吕小康．从工具到范式：假设检验争议的知识社会学反思[J]．社会，2014，34(06)：216-236.
[5]Wasserstein R L，Lazar N A．The ASA Statement on P-values：Context，Process，and Purpose[J]．The American Statistician，2016，70(2)：129-133.
[6]Baker M．Statisticians Issue Warning over Misuse of p Values[J]．Nature News，2016， 531(7593)：151.
[7]郝丽，刘乐平，申亚飞．统计显著性：一个被误读的 P 值——基于美国统计学会的声明[J].统计与信息论坛，2016，31(12)：3-10.
[8]Lindley D V．The Analysis of Experimental Data：The Appreciation of Tea and Wine[J]．Teaching Statistics，1993，15(1)：22-25.
[9]Nuzzo R．Scientific Method：Statistical Errors[J]．Nature News，2014，506(7487)：150.
[10]Newman M C．“What Exactly Are You Inferring?” A Closer Look at Hypothesis Testing[J]．Environmental Toxicology and Chemistry：An International Journal，2008，27(5)：1013-1019.
[11]Hubbard R，Bayarri M J．Confusion Over Measures of Evidence (p’s) Versus Errors (α’s) in Classical Statistical Testing[J]．The American Statistician，2003，57(3)：171-178.
[12]Fisher R A．The Design of Experiments[M]．London：Oliver and Boyd，1935.
[13]吕小康．Fisher 与 Neyman-Pearson 的分歧与心理统计中的假设检验争议[J]．心理科学，2012，35(06)：1502-1506.
[14]Gill，J．The Insignificance of Null Hypothesis Significance Testing[J]．Political Research Quarterly，1999，52(3)：647-674.
[15]Gigerenzer G．Mindless statistics[J]．The Journal of Socio-Economics，2004，33(5)：587-606.
[16]Neyman J．Frequentist Probability and Frequentist Statistics[J]．Synthese，1977：97-131.
[17]Szucs D，Ioannidis J．When Null Hypothesis Significance Testing is Unsuitable for Research：A Reassessment[J]．Frontiers in Human Neuroscience，2017，11：390.
[18]权朝鲁．效果量的意义及测定方法[J]．心理学探新，2003(02)：39-44.
[19]Cohen J．The Statistical Power of Abnormal-social Psychological Research：A Review[J]．The Journal of Abnormal and Social Psychology，1962，65(3)：145.
[20]Sedlmeier P，Gigerenzer G．Do Studies of Statistical Power Have an Effect on the Power of Studies?[J]．Psychological Bulletin，1989，105(2)：309-316.
[21]Perezgonzalez J D．The Meaning of Significance in Data Testing[J]．Frontiers in Psychology，2015，6：1293.
[22]Perezgonzalez，Jose D．Fisher，Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing[J]．Front Psychol，2015，6：223.
[23]吴喜之．统计学：从数据到结论[M]．第 4 版．北京：中国统计出版社，2013.
[24]Patriota A G．On Some Assumptions of the Null Hypothesis Statistical Testing[J]．Educational & Psychological Measurement，2016，77(3)：507－528.
[25]Chang M．What Constitutes Science and Scientific Evidence：Roles of Null Hypothesis Testing[J]．Educational and Psychological Measurement，2016，77(3)：475－488.
[26]Häggström O．The Need for Nuance in the Null Hypothesis Significance Testing Debate[J]．Educational and Psychological Measurement，2016，77(4)：616－630.
[27]Sesardic N．Sudden Infant Death or Murder? A Royal Confusion About Probabilities[J]．The British

Journal for the Philosophy of Science, 2007, 58(2): 299－329.
[28]Hubbard R, Bayarri M J. Confusion Over Measures of Evidence (p’s) Versus Errors (α’s) in Classical Statistical Testing[J]. The American Statistician, 2003, 57(3): 171－178.
[29]茆诗松，周纪芗. 概率论与数理统计[M]. 第1版. 中国统计出版社，1996.
[30]陈希孺. 数理统计学教程[M]. 第1版. 中国科技大学出版社，2009.
[31]Pollard P, Richardson J T. On the Probability of Making Type I Errors[J]. Psychological Bulletin, 1987, 102(1): 159－163.
[32]Wagenmakers E J. A Practical Solution to the Pervasive Problems of P Values[J]. Psychonomic Bulletin & Review, 2007, 14(5): 779-804.
[33]Schneider, Jesper W. Null Hypothesis Significance Tests. A Mix-up of Two Different Theories: the Basis for Widespread Confusion and Numerous Misinterpretations[J]. Scientometrics, 2015, 102(1): 411-432.
[34]Jeffreys H. Theory of probability[M]. Oxford: OxfordUniversityPress, 1961.
[35]Rosenthal R, Gaito J. The Interpretation of Levels of Significance By Psychological Researchers[J]. The Journal of Psychology, 1963, 55(1): 33－38.
[36]Nelson N, Rosenthal R, Rosnow R L. Interpretation of Significance Levels and Effect Sizes By Psychological Researchers[J]. American Psychologist, 1986, 41(11): 1299－1301.
[37]Peto R, Pike M C, Armitage P, et al. Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. I. Introduction and Design[J]. British Journal of Cancer, 1976, 34(6): 585－612.
[38]Bakan D. The Test of Significance in Psychological Research[J]. Psychological Bulletin, 1966, 66(6): 423－437.
[39]Fisher R A, Statistical Methods for Research Workers[M]. London:Oliver And Boyd, 1950.
[40]Krueger J. Null Hypothesis Significance Testing. On The Survival of A Flawed Method[J]. Am Psychol, 2001, 56(1): 16-26.
[41]仲晓波. 关于假设检验的争议:问题的澄清与解决[J]. 心理科学进展，2016，24(10): 1670-1676.